Binaural Cue Coding—Part I: Psychoacoustic Fundamentals and Design Principles

Frank Baumgarte and Christof Faller

***Abstract*—Binaural Cue Coding (BCC) is a method for multichannel spatial rendering based on one down-mixed audio channel and BCC side information. The BCC side information has a low data rate and it is derived from the multichannel encoder input signal. A natural application of BCC is multichannel audio data rate reduction since only a single down-mixed audio channel needs to be transmitted. An alternative BCC scheme for efficient joint transmission of independent source signals supports flexible spatial rendering at the decoder.**

This paper (Part I) discusses the most relevant binaural perception phenomena exploited by BCC. Based on that, it presents a psychoacoustically motivated approach for designing a BCC analyzer and synthesizer. This leads to a reference implementation for analysis and synthesis of stereophonic audio signals based on a Cochlear Filter Bank. BCC synthesizer implementations based on the FFT are presented as low-complexity alternatives. A subjective audio quality assessment of these implementations shows the robust performance of BCC for critical speech and audio material. Moreover, the results suggest that the performance given by the reference synthesizer is not significantly compromised when using a low-complexity FFT-based synthesizer. The companion paper (Part II) generalizes BCC analysis and synthesis for multichannel audio and proposes complete BCC schemes including quantization and coding. Part II also describes an alternative BCC scheme with flexible rendering capability at the decoder and proposes several applications for both BCC schemes.

***Index Terms*—Audio coding, auditory filter bank, auditory scene synthesis, binaural source localization, coding of binaural spatial cues, spatial rendering.**

I. Introduction

THE data rate of traditional subband audio coders, such as AAC [1] and PAC [2], scales with the number of audio channels. If the channels are compressed independently, the data rate grows proportionally to the number of channels. Joint-channel coding techniques, such as "Sum-Difference Coding" [3], "Intensity Stereo Coding" (ISC) [4], and "Inter-Channel Prediction" [5] can reduce this growth rate. However, the resulting data rate for conventional stereophonic[1] material is still considerably higher than needed for representing the corresponding mono audio signal. Thus, the trade-off between audio bandwidth, coding artifacts, and number of channels typically dictates to use only one mono channel if the target data rate falls below a certain threshold.

Manuscript received May 25, 2002; revised August 6, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gerald Schuller.

The authors are with the Media Signal Processing Research Department, Agere Systems, Allentown, PA 18109 USA (e-mail: fb@agere.com; cfaller@agere.com).

Digital Object Identifier 10.1109/TSA.2003.818109

[1]The term "stereo" or "stereophonic" always refers to two-channel stereophonics only.

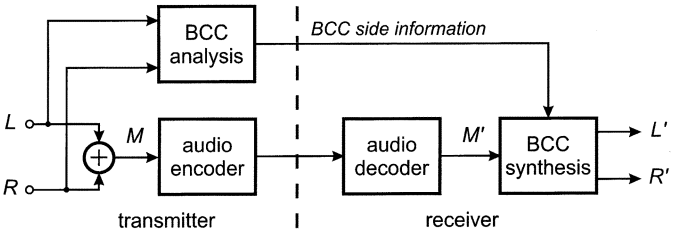

Fig. 1. Generic BCC scheme enhancing a mono audio coder to stereo.

Binaural Cue Coding (BCC) offers a solution for providing multichannel audio at low and very low data rates. BCC was introduced in [6]–[12]. Selected results from these publications are included here to provide a complete overview. A basic scheme for coding a stereophonic signal with BCC is shown in Fig. 1 as an example. Such a scheme is referred to as BCC for Natural Rendering, aka type II BCC. In the transmitter, a BCC analyzer extracts binaural spatial cues from the original stereophonic signal, L and R. The stereophonic signal is down-mixed to mono and compressed by a suitable audio encoder. In the receiver, the mono audio signal is decoded. The BCC synthesizer reconstructs the spatial image by restoring spatial localization cues when it generates the stereophonic output from the mono signal. This scheme is closely related to full-bandwidth ISC. The reasons why BCC can be applied to the full audio bandwidth while ISC has the drawback of being only suitable for the mid-to-high frequency range are discussed in [9].

The BCC side information contains the spatial localization cues and can be transmitted with a rate of only a few kb/s. The independent representation of the information for reproducing the spatial image in the BCC scheme allows to control spatial image distortions or modifications separately from the mono audio coding scheme. Since the BCC analysis and synthesis are separated from the mono audio coder, existing mono audio or speech codecs can be enhanced for multichannel coding with BCC. It is interesting to note that this BCC scheme prevents practically all binaural unmasking effects that otherwise have to be considered in conventional multichannel coders [13]. This property arises from the fact that both, a virtual sound source (phantom source) and the associated quantization noise created by the audio coder, will be localized in the same spatial direction. Thus, a condition that implies a binaural masking level difference (BMLD) [14] does not occur.

This paper focuses on psychoacoustic considerations in the system design of BCC and the perceptual implications of dif-

ferent implementations. Section II summarizes basic psychoacoustic facts of spatial perception explored by BCC. We point out existing binaural perception models as a first approach to extract the spatial cues. From psychoacoustic considerations, an ideal BCC analysis/synthesis scheme can be formulated. The corresponding BCC implementation is used as a reference for stereophonic natural rendering. Less complex implementations are derived to reduce computational costs. Implementations of the analysis and synthesis are described in Section III. Their performance is evaluated in Section IV. This includes an example of BCC spatial cue estimation data and various psychoacoustic test results for the reproduced spatial image and audio quality. In Section V we conclude our findings and discuss future work.

Part II [15] addresses quantization and coding of the BCC side information, which is excluded here. A second scheme referred to as BCC for Flexible Rendering (aka type I BCC) is introduced in Part II and implementation details of both schemes, for natural and flexible rendering, are given. Generalizations include generic multichannel audio and synthesis of an enhanced set of spatial cues.

II. Psychoacoustic Considerations for BCC Design

The derived audio quality of the BCC scheme relies heavily on the ability to synthesize the proper auditory spatial cues in the stereophonic signal in the BCC synthesis block of Fig. 1. From psychoacoustics it is well-known how to synthesize two-channel signals in order to create the illusion of a sound source (phantom source) at a certain position [14]. These techniques allow control of direction (azimuth and elevation) as well as distance. Furthermore, the source width and blur can be manipulated. In sound engineering, techniques have been developed to create reverberation, envelopment, depth, and other spatial attributes of sound [16]. However, most of the above-mentioned technologies create identical sound attributes for auditory objects of a one-channel input, e.g. all physical sound sources mixed in one microphone signal. For example, amplitude panning [17] with mixing consoles can change the apparent azimuth, but it cannot control independently the azimuth of a violin and a piano unless the instruments are recorded separately. Thus, existing techniques in sound engineering and psychoacoustics are not sufficient to spatially separate sound sources contained in a mono signal, which is essential for the BCC synthesizer when recreating a spatial image. This problem will be addressed after discussing the properties of binaural spatial cues that evoke the spatial image.

Ideally, the synthesizer produces an audio signal that evokes the same binaural spatial cues at both ears as the stereophonic original. In the following we consider only the most important binaural cues [14], "interaural level difference" (ILD), "interaural time difference" (ITD), and "interaural correlation" (IC), since an exhaustive discussion of all possible spatial cues cannot be given here. Depending on the playback scenario, however, we only have limited control over the binaural cues since the acoustical transfer functions (ATFs) from the transducers to both ears are not precisely known in advance and can only be roughly estimated for most applications. However, we can ignore the ATFs here if we assume that their impact on the spatial image is similar for the playback of the original and synthesized audio. This assumption is supported by subjective test results in Section IV. Consequently, we do not generally optimize the reproduction of binaural cues at both ears but we optimize the recreation of spatial cues contained in the transducer signals. To distinguish both cases, we introduce the terms "inter-channel level difference" (ICLD), "inter-channel time difference" (ICTD), and "inter-channel correlation" (ICC) which refer to the transducer signals, i.e. the customary audio signal. For headphone playback, the ICLD, ICTD, and ICC are virtually identical to the ILD, ITD and IC, respectively. ICLD and ICTD determine the lateralization of strongly correlated signals. A decreasing ICC is perceived as increasing source width until the phantom source splits into two sources, one at the left and the right side. Decreasing ICC can also enlarge the apparent distance in case of loudspeaker playback [14].

To point out how BCC can recreate a given spatial image, we look at a stereophonic signal composed of two phantom sources at different locations as an example. For this case, we assume that all signal components of either one phantom source create a consistent ILD and ITD at the listener's ears. Signal components are areas in the time-frequency plane with significant energy. However, the cues of one source are corrupted by the cues of the other in areas of the time-frequency plane where both sources have significant energy. Nevertheless, the auditory system has the ability to perceptually segregate the sources with fascinating accuracy. This ability can be explained by perceptual grouping mechanisms applied to the sound components, that only partially rely on spatial cues [18]. For example, "similar" spectral components are often fused to one auditory object, independent of the consistency of the spatial cues. Even if contradicting spatial cues are present, the spatial image often appears robust. Depending on frequency and audio signal content, certain cues dominate the determination of the auditory spatial image [19]. For instance, the well-established Duplex Theory [19], [20] states that ILD cues are most salient above ca. 1.5 kHz. At lower frequencies ITD cues are more relevant [21].

These perceptual factors contribute to the robustness of the audio quality of BCC with respect to a nonideal spatial cue analysis and synthesis with a low-complexity filter bank as shown in the results (Section IV). This robustness is also observed when applying coarse quantization to the localization cues as described in Part II [15]. The consequences of auditory scene analysis for BCC are discussed more detailed in [11].

In case of loudspeaker playback, the spatial cues present at the listener's ears are a function of the cues present in the transducer signals and the ATFs from the transducers to the ear entrances. Assuming the simplest acoustical condition, i.e. the free field, we observe from measurements and simulations that the ICLD is translated into an ITD at the listener's ears for frequencies below ca. 1.5 kHz [11]. This phenomenon is a consequence of the acoustical properties of the listener's head in the sound field [14]. It is replicated by our simulation results in [11]. Based on this property and considering the Duplex Theory, the most salient localization cues can be provided with loudspeaker playback, even if ICTDs are ignored. The traditional two-channel stereo system ("Blumlein" stereo) and associated audio format relies on these conditions. However, if the loudspeakers are located in a reverberant environment, spatial localization cues can

be considerably influenced or even dominated by room acoustics. The effects of reverberation are beyond the scope of this paper. To simplify the following discussion we exclude the ATFs by assuming headphone playback. Still, the conclusions are applicable to loudspeaker playback if the ATFs are taken into account. For headphone reproduction it should be noted that the ICTDs at low frequencies are salient cues according to the Duplex Theory.

For the BCC analysis, we need to explore knowledge about binaural processing in order to closely approximate the internal cues as they occur in the auditory system. On this basis, the synthesis can be designed such that it generates an output that evokes similar binaural cues. For most purposes, it is trivial to generalize the two-channel case, treated here, to the generic multichannel case [15].

A straight-forward design of an analyzer can be based on existing binaural perception models. Most of the models aim at predicting binaural detection thresholds, some also include localization in terms of azimuth and lateralization. A review of suitable models is given in [22]. A further advanced model described in [23] is particularly versatile and represents the state of the art since it successfully reproduces a wide variety of binaural perception phenomena. A preprocessing block that covers the signal processing of the peripheral ear, including the cochlea, is common to most binaural models. It includes a spectral decomposition into critical bands and usually the rectification and low-pass filtering associated with the inner hair cells. The output resembles the firing rate of the auditory nerve. After the preprocessing, many models apply a correlation analysis to corresponding outputs of the left and right peripheral ear models as the first stage of the binaural processor. This approach is based on the "coincidence counter hypotheses" discussed in [23]. The location of the correlation maximum indicates the ITD. The underlying power estimation for the correlation analysis can be used to derive the ILD. Furthermore, the correlation maximum value indicates the perceived image width. An alternative modeling approach for the binaural processing is based on the "equalization/cancellation" theory [23]. This type of model is in many respects equivalent to the correlation based approach. However, the signal processing of this model type does not directly support an estimation of the correlation which is desirable for the BCC synthesis.

Usually, binaural detection models are designed and verified based on reference stimuli consisting of a single discrete sound source. We assume that these modeling approaches are applicable as well for the binaural cue estimation of multiple simultaneous sources, since the models are based on a generic peripheral preprocessing stage for deriving the cues, which does not imply a source number limitation. This assumption is complemented by a second pivotal assumption, stating that the auditory spatial image can be reconstructed from a mono signal by restoring the estimated binaural cues according to the scheme in Fig. 1. This second assumption implies that a traditional computational auditory scene analysis is not necessary since we only deal with binaural cues without assigning them to any specific sound source. Given this framework, the restoration of estimated binaural cues in the BCC synthesizer is ideally done by a processing scheme implementing the inverse binaural model used for the cue estimation. Such a scheme for analysis and synthesis of binaural cues is presented in Section III-A.

Even if perfect reconstruction of the spatial image would be possible, the down-mixing to a mono signal usually results in a loss of perceptually relevant information. For instance, if left and right components are 180° out of phase and of equal level, they cancel each other and will not be recoverable. However, for playback compatibility reasons, most existing stereophonic recordings are "mono compatible" in the sense that a maximum of audio content is preserved after the down-mixing. Moreover, advanced down-mixing techniques aiming at preserving the spectral energy distribution or loudness can be applied to circumvent coloration effects. Such a technique is described in Part II [15].

Important parameters of the analysis and synthesis schemes are the effective "critical" bandwidth and time resolution of binaural hearing, especially for localization tasks. Psychoacoustic masking experiments show that the critical bandwidths in binaural detection tasks within a range of center frequencies between 200 and 1000 Hz are equal [24] or up to about 50% larger than monaural critical bands depending on the type of experiment [25], [26]. A suitable numerical definition for monaural critical bandwidths is given in [27]. As opposed to the bandwidths, the binaural time resolution is significantly different from the monaural time resolution. Different detection experiments at 500 Hz reveal monaural time constants in a range of 2–26 ms and binaural time constants of 33–189 ms [25]. The experiments were done with a pure tone masked by noise presented monaurally or binaurally out of phase for the binaural case. The experimentally derived numbers provide a guideline for interpreting the psychoacoustic experiments presented in Section IV because they are considered to be also relevant for the binaural cue estimation and synthesis.

III. BCC Analysis and Synthesis

A reasonable approach to realize a BCC analyzer and synthesizer for natural rendering consists of utilizing knowledge from existing binaural models. This approach is adopted here by using a Cochlear Filter Bank (CFB) for the analysis with a time and frequency response similar to the human inner ear. The synthesis makes use of a corresponding inverse CFB. This is an ideal scheme with respect to the design goals formulated and in the sense of the assumptions made in Section II. It will be used as a reference for other schemes described, based on the Fast Fourier Transform (FFT). The main motivation to use filter banks different from the CFB is a reduction of computational complexity. All schemes presented here are limited to ICLD synthesis and they were partially presented in [7], [11]. Part II includes ICTD and ICC synthesis in enhanced schemes.

A. Analysis and Synthesis Based on a Cochlear Filter Bank

Suitable binaural models [22], [23] apply a filter bank as first processing stage that has similar properties as the frequency decomposition found in the inner ear. A filter bank with equivalent properties but with a particularly efficient implementation is given in [28]. This CFB is used for the BCC analysis. The corresponding inverse CFB is described here and is applied for

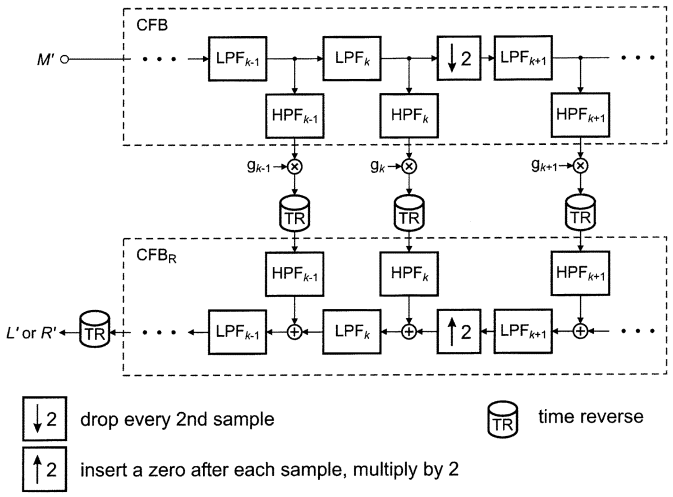

Fig. 2. Structure of the forward Cochlear Filter Bank (CFB) and its inverse. The inverse CFB includes time reversals (TRs) and the reverse CFB $(\mathrm{CFB_R})$. The low-pass filters (LPFs) and high-pass filters (HPFs) are identical for the CFB and $\mathrm{CFB_R}$. The coefficients g_k are needed for equalization and optionally for ICLD synthesis.

the BCC synthesis. Fig. 2 shows a block diagram of the forward and inverse filter-bank structure.

The forward structure [28] consists of a low-pass filter (LPF) cascade with down-samplers. Each low-pass output is processed by a high-pass filter (HPF) to generate the band-pass signals at the CFB output. These outputs represent "critical band" signals that overlap spectrally. The input audio signal can only be approximately reconstructed from the critical-band signals by applying the inverse filter bank. The inverse CFB includes the reverse structure $(\mathrm{CFB_R})$ and time reversal operations. It is derived by reversing the signal flow, replacing down-samplers by up-samplers, and by time reversing the filter impulse responses of the forward CFB. The time reversal of the impulse responses is substituted by applying time reversal to the input and output signals of the inverse CFB. This substitution allows a less complex implementation than the reversal of the IIR-type filter responses. For signals that are not time limited, the time reversal can be implemented by block-wise processing with temporal overlap, as described in [29]. The signal processing for the experiments reported in this paper was done by time reversing all $(\mathrm{CFB_R})$ full-length band-pass input signals at once and time reversal of the output signal after filtering as outlined in Fig. 2. The gain factors, g_k, for the band-pass signals are needed to compensate for the energy increase due to the band overlap of neighboring filters. Furthermore, these factors include the level modification for ICLD synthesis.

For the analysis, only the forward CFB is necessary and complemented by a simple inner hair cell (IHC) model in each band as shown in Fig. 3. The IHC model includes a half-wave rectifier and low-pass filter (LPF). The LPF is composed of two identical cascaded first-order filters with a cutoff frequency of $f_{\mathrm{c,IHC}}$ according to (1). The center frequency of the CFB band in Hz is denoted f_{CFB}. The parameter f_0 is chosen to be $f_0 = 300$ Hz. The CFB and inner hair cell model parameters are adopted from [28]

$$f_{\mathrm{c,IHC}} = \begin{cases} f_0 & \text{if } f_{\mathrm{CFB}} < f_0 \\ f_0 \left(\frac{f_{\mathrm{CFB}}}{f_0}\right)^{0.25} & \text{if } f_{\mathrm{CFB}} \geq f_0 \end{cases}. \quad (1)$$

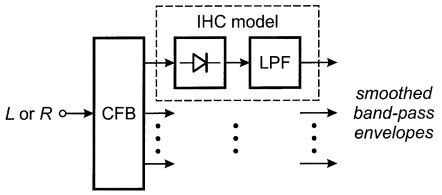

Fig. 3. Model of peripheral auditory processing including Cochlear Filter Bank (CFB) and inner hair cell model (IHC) with low-pass filter (LPF).

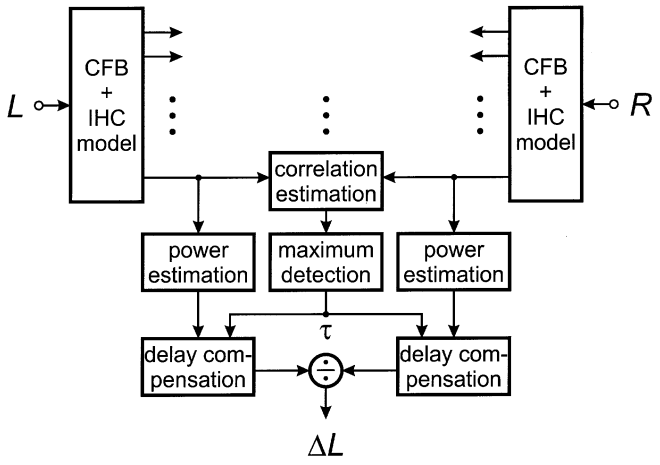

Fig. 4. Block diagram of CFB-based BCC analyzer.

A smooth envelope is derived at the outputs of the CFB bands at medium and high center frequencies. For simplicity, all outputs are referred to as band-pass envelopes even though at low center frequencies the waveform is still present. The CFB outputs have a maximum delay of 10 ms that decreases with increasing center frequency. The delay of all bands is equalized by adding the necessary delay in each band. This simplifies the application of the estimated cues if the synthesizer is based on a uniform filter bank with constant delay.

Fig. 4 shows the derivation of ICTDs, τ, and ICLDs, ΔL, in each CFB band for a pair of channels (e.g. left and right channel). The estimation of the ICTDs is based on a normalized cross-correlation measure $\hat{\gamma}_{xy}$, the ICC. The ICC is derived from a cross-correlation estimate $\hat{\phi}_{xy}$ normalized by the auto-correlation estimates $\hat{\phi}_{xx}$ and $\hat{\phi}_{yy}$ of the signals in both channels L and R according to (2) and (3). The time shift m is expressed as number of sampling intervals. The index of the current sampling interval (time index) is i

$$\hat{\gamma}_{xy}(m,i) = \frac{\hat{\phi}_{xy}(m,i)}{\beta(m,i)} \quad (2)$$

$$\beta^2(m,i) = \begin{cases} \hat{\phi}_{xx}(0,i-m)\hat{\phi}_{yy}(0,i) & \text{if } m \geq 0 \\ \hat{\phi}_{xx}(0,i)\hat{\phi}_{yy}(0,i+m) & \text{if } m < 0 \end{cases}. \quad (3)$$

The mean value is removed from the smoothed band-pass envelopes. The result is denoted x and y corresponding to the input audio channels L and R, respectively, in one representative CFB band. The time shift between the envelopes x and y is m. The cross-correlation function is estimated recursively using (4) and (5)

$$\hat{\phi}_{xy}(m,i) = w\hat{\phi}_{xy}(m,i-1) + [1-w]\delta(m,i) \quad (4)$$

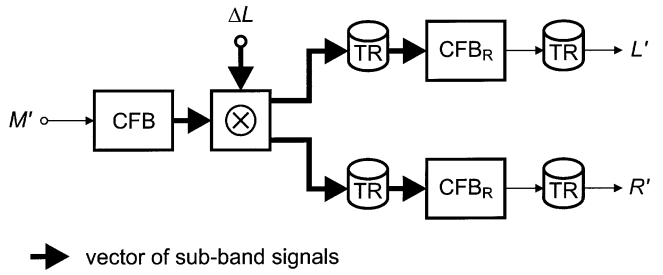

Fig. 5. Block diagram of CFB-based BCC synthesizer (see Fig. 2 for legend).

$$\delta(m,i) = \begin{cases} x(i-m)y(i) & \text{if } m \geq 0 \\ x(i)y(i+m) & \text{if } m < 0 \end{cases}. \quad (5)$$

The time constant of the exponential estimation window is determined by w. It was adjusted such that the estimated ICTDs and ICLDs based on the ICC are able to track changes in the input ICC fast enough while maintaining a reasonably stable result for a stationary ICLD and ICTD for natural sound sources like speech or vocals. A good compromise is achieved with $w = 0.998$. This value corresponds to a cutoff frequency of about 10 Hz for a sampling rate of $f_s = 32$ kHz if (4) is interpreted as recursive low-pass filtering with the filter coefficient w.

The auto-correlation in (3) used for the normalization is estimated according to (6) and (7). The same factor w as in (4) must be used here to maintain the desired ICC range between -1 and 1.

$$\hat{\phi}_{xx}(0,i-m) = w\hat{\phi}_{xx}(0,i-m-1) + [1-w]x^2(i-m) \quad (6)$$

$$\hat{\phi}_{yy}(0,i+m) = w\hat{\phi}_{yy}(0,i+m-1) + [1-w]y^2(i+m). \quad (7)$$

The ICTD is estimated by locating the maximum of the ICC $\hat{\gamma}_{xy}(m,i)$ with respect to m. If the maximum is located at $m = m_{\max}$, the ICTD is $\tau = m_{\max}/f_s$. The ICLD estimation is based on the ratio of the estimated band powers. The power estimation uses a recursive low-pass filter applied to the squared inner hair cell model outputs. The filter cutoff frequency is about 50 Hz. The ICLD, ΔL, is the ratio of the ICTD-compensated power estimates from both channels and converted to the logarithmic (dB) domain.

The ICC is computed for a limited symmetrical range of delays m with respect to zero delay because auditory localization based on ITDs "saturates" at the extreme left or right of the auditory space for delays larger than approximately 1 ms. However, we currently use a delay range of ± 1.6 ms to get an improved ICLD estimate if larger ICTDs are present.

An overview of the CFB-based synthesis scheme is given in Fig. 5. The mono audio signal is decomposed into critical bands by the forward CFB. The estimated cues are applied to the band-pass signals. Currently, only ICLD synthesis is implemented. This is done by modifying the gains, g_k, in Fig. 2 according to the estimated ICLDs, ΔL, for the left and right channel (see Part II [15] for details). In principle, the synthesis of ICTD and ICC modifications can also be done in the band-pass signal domain.

B. Analysis and Synthesis Based on the FFT

The computational complexity of the CFB-based BCC implementation is relatively high when compared with other coding algorithms, e.g., Intensity Stereo Coding. Since the filter bank complexity (as given in [28]) dominates the BCC analysis and synthesis costs, lower complexity filter banks are of interest as a replacement for the CFB. Basically, all modulated filter banks that can be implemented with an FFT-type fast algorithm are candidates. In particular, nonuniform modulated filter banks [30] seem attractive since they can approximate the auditory frequency resolution to a certain extent. A first study [9] that compares BCC synthesis schemes based on the Modified Discrete Cosine Transform (MDCT) [31] and the FFT concludes that the FFT has superior performance for BCC. The FFT can be interpreted as a modulated filter bank. Its spectral representation is well suited for BCC since it supports simple ICTD synthesis. The estimated potential complexity (instructions/second) reduction by using the FFT instead of the CFB can reach two orders of magnitude. The details of FFT-based BCC analysis and synthesis algorithms supporting ICLD and ICTD are presented in [15]. Hence, we give here only a brief overview of FFT-based BCC and focus on the comparison with the CFB-based reference BCC.

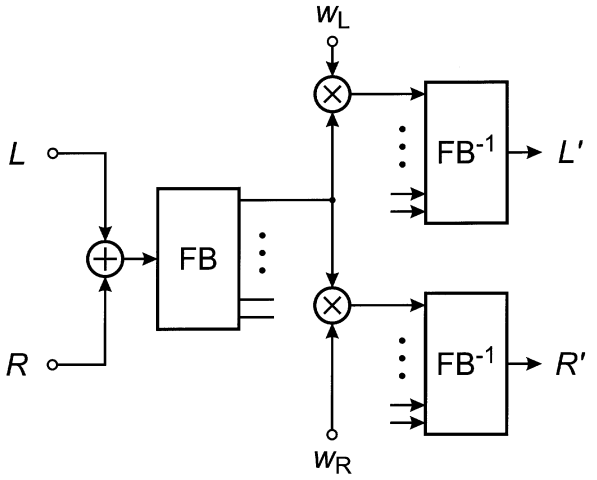

Fig. 6. Generic filter-bank-based BCC synthesis scheme for inter-channel level differences (ICLDs).

For an FFT-based analysis, a standard block-wise short-time FFT decomposition of the audio signal with 50% overlapping windows is used. The FFT spectrum is divided into nonoverlapping partitions, each representing an auditory "critical band." The partition bandwidth used is 2 ERB (Equivalent Rectangular Bandwidth) [27]. In contrast to the CFB-based analyzer, IHC models are not included to minimize complexity. The ICLDs are estimated by calculating the power ratio of the corresponding partitions of L and R. At low frequencies, the ICTD of each partition corresponds to the phase difference between a channel pair. At medium and high frequencies, ICTDs are derived from the slope of the phase difference between L and R versus frequency which indicates the group (envelope) delay. This is motivated by the low-pass effect of the IHCs, that creates the corresponding band-pass signal envelopes. Only one averaged ICLD and one ICTD value per partition is provided to the synthesis.

The FFT-based synthesis used in the experiments reported here performs the same spectral decomposition as the analysis described above. Fig. 6 shows the synthesis of ICLDs in the frequency domain by modifying the magnitude spectrum. The ratio

of the weighting factors w_L and w_R is equal to the ICLD, ΔL. The absolute value of the factors is determined by demanding that the subband power sum of the left and right channel is equal to the mono signal. ICTDs, τ, can be synthesized by modifying the phase-difference spectrum between the two channels as described in Part II [15] to create the desired phase differences and slopes of phase differences. The inter-channel correlation (ICC) of the synthesizer output can be reduced by additional modifications of ICLDs and/or ICTDs. Such modifications aim to preserve the average ICLD and ICTD in each partition while reducing the ICC by complementary changes of ICLDs and/or ICTDs within the partition, for example [15]. The modified spectra for the left and right channel are finally transformed back to the time domain.

The CFB-based reference synthesis allows to introduce slowly time-varying ICLDs without creating audible artifacts, such as aliasing in the time or frequency domain. This desirable property is a consequence of the high attenuation of the CFB filters at the Nyquist frequency, which is more than 100 dB for most filters. In contrast, many uniform filter banks use aliasing cancelation to achieve high stop-band attenuation and critical down-sampling. However, aliasing cancelation is reduced if the band-pass signals are modified, e.g. by ICLDs. Specifically, the FFT-based BCC scheme can produce different types of distortions. These distortions include frequency and time-domain aliasing, and "blocking" artifacts. Blocking artifacts can occur if the spatial cues vary over time and the inverse FFT is applied without a smoothing overlap for consecutive audio signal blocks. Overlapping windows that provide a "fade-in" and "fade-out" transition avoid this kind of artifact. For the experiments reported here, we multiplied the input audio block with a sine window before applying the FFT. This window is equal to the first half cycle of a sinusoid. After the synthesis with the inverse FFT, we use the same sine window before performing the overlap-add with the previous block. The effect of applying the window twice is to impose a squared sine window (Hann window) on the synthesized data block. With subsequent overlap-add using 50% overlap, the effective squared sine windows of subsequent blocks add up to a constant value of 1 as desired.

Frequency-domain aliasing distortions can occur if the Fourier spectrum is modified by the restoration of spatial cues before applying the inverse FFT. In particular, the aliasing components are not fully canceled by the inverse FFT if spectral components of neighboring bands are modified differently. Audible aliasing distortions can be avoided by limiting the amount of variation permitted for spectral modification of neighboring FFT bands. Moreover, frequency-domain aliasing can be reduced by increasing the oversampling rate of the FFT, i.e. increasing the window overlap size at constant block length. A more detailed discussion of aliasing effects in the context of BCC can be found in [9].

Time-domain aliasing distortions can be created when the Fourier spectrum is modified by synthesizing spatial cues. This is particularly a problem if ICTDs are applied. The introduced time delay generally results in a circular time shift of the synthesized block. The time aliasing created by the circular shift can be avoided by using zero-padded windows. Thus, no time-domain aliasing will occur up to a certain maximum time delay. A second problem arises for the ICTD synthesis if a synthesis window is applied as described above. After creating a time delay, the synthesis time window is misaligned in time with the "delayed" analysis window. To circumvent this misalignment, a rectangular synthesis window, i.e., no synthesis window, is used for ICTD synthesis. Consequently, the analysis window is applied twice. That effectively results in a Hann window for the analysis FFT. This window combination is specified and applied in Part II [15].

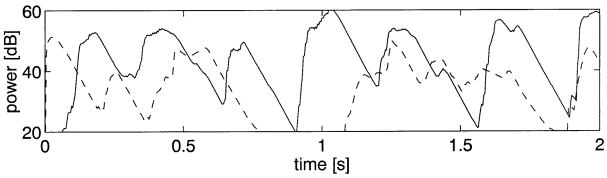

Fig. 7. Estimated signal power of the two talkers in the frequency band centered at 1008 Hz. Talker 1: dashed, talker 2: solid.

IV. Results

This section presents methods to assess the performance of different BCC implementations and experimental results. The purpose of these assessments is the validation that BCC can provide suitable audio quality for the targeted applications. Moreover, the quality degradation of low-complexity synthesizer implementations with respect to the reference CFB-based scheme is evaluated. These results were partly presented in [7], [9], [11]. Subjective tests of low-complexity FFT-based analyzers are not included here (see Part II [15]). All subjective tests were carried out using high quality loudspeakers (B&W Nautilus 802) since loudspeaker playback is assumed to be a more common playback scenario than headphones for the anticipated applications. For loudspeaker playback, we assume that the synthesis of ICTDs is not necessary for the spatial image reproduction as discussed in Section II. Furthermore, for the experiments reported here we neglect the correlation cues and concentrate only on ICLD synthesis.

The first experiment (Section IV-A) illustrates the estimation accuracy of the CFB-based analyzer. The second experiment is concerned with the subjective quality obtained from the combination of this analyzer with an FFT-based synthesizer. The third experiment (Section IV-B) compares the perceived audio quality obtained with the same analyzer combined with the CFB-based synthesizer and FFT-based synthesizers of different FFT size.

A. CFB-Based Analysis Combined With FFT-Based Synthesis

The performance of the CFB-based spatial cue estimation described in Section III-A is illustrated for speech signals. For that purpose different stereophonic signals were created by amplitude and/or time-delay panning and superposition of two separate talkers. The estimated inter-channel cues are compared to the "ideal" cues applied. Fig. 7 shows the estimated power of the two separate one-channel talker signals at the output of the filter with 1008 Hz center frequency. This filter has a 3-dB bandwidth of approximately 100 Hz [28]. The panning of the two

TABLE I
PARAMETERS OF SYNTHESIZED STEREOPHONIC SIGNALS A, B, C

label	*talker 1*		*talker 2*	
	τ_{ref} [ms]	ΔL_{ref} [dB]	τ_{ref} [ms]	ΔL_{ref} [dB]
A	0	10	0	−10
B	0.6	0	−0.6	0
C	0.6	10	−0.6	−10

mono signals was done according to Table I to create a stereophonic signal with ICLDs, ΔL_{ref}, only (A), with ICTDs, τ_{ref}, only (B), and a combination of both (C). The estimated ICTDs in Fig. 8 for B and C show a quasi instantaneous transition between the ICTDs of both talkers since the correlation function can have two maxima corresponding to the two delays. Since the larger maximum is chosen for the ICTD estimation, basically a switching between both values occurs. The ICTD estimate for A is almost ideally zero except for a few single values. Fig. 9 shows the estimated ICLDs for all three signals. It appears almost identical for A and C as expected. Due to the overlap of both talkers the ICLD gradually changes between the ICLD of one talker to the ICLD of the other. For B the estimate is close to the applied ICLD of zero as desired.

In the second experiment, we attempted to assess the perceived quality of the reproduced spatial image generated by an FFT-based BCC synthesizer. Synthesized stereophonic signals with two or three discrete phantom sources were used as audio material. The phantom sources were created by amplitude panning which imposes an ICLD, $\Delta L_{\mathrm{ref},n}$, onto the time-aligned (zero ICTD) signals of the n-th source in the two channels.

The choice of synthesized signals as opposed to natural stereophonic recordings is motivated by having better control over the spatial image and strictly defined parameters for creating the image. A further advantage of the synthesized signals is the absence of reverberation and reflections from other spatial directions than the sound source direction. Based on that, the subjective assessment is better controlled and the perceptual task is simplified. However, these restrictions must be dropped if a performance assessment for more generic signals is required.

Four different categories of signal sources were used: single talkers (D), solo vocals (E), keyboard instruments (F), and percussive instruments (G). Four reference signals $(\mathrm{D2}\ldots\mathrm{G2})$ were generated by mixing two sources of the same category with ICLDs of 10 dB and −10 dB. Another four reference signals $(\mathrm{D3}\ldots\mathrm{G3})$ were generated by mixing three sources of the same category with ICLDs of 10 dB, −10 dB, and 0 dB. Sources of the same category were mixed because they are most likely to have an impact on each other's phantom image due to their similar time-frequency characteristics. The audio excerpts were selected from a collection of critical signals with the objective of having different types of content and the most critical material for ICLD imaging in the test.

To facilitate the evaluation of the listening test results, two types of anchor signals were also presented in the test. For the first type, the ICLDs are sinusoidally modulated over time with a frequency of 0.5 Hz to create moving phantom sources. The ICLD varies between 10 dB and 5 dB instead of 10 dB in the reference, between −10 dB and −5 dB instead of −10 dB, and between −2.5 dB and 2.5 dB instead of 0 dB. The second type adds localization blur by modifying the ICLDs of every other partition. The modified partitions have 5 dB ICLD instead of 10 dB in the reference and −5 dB ICLD instead of −10 dB. The ICLD of 0 dB in the reference is replaced by alternating the ICLDs in the partitions between −2.5 dB and 2.5 dB.

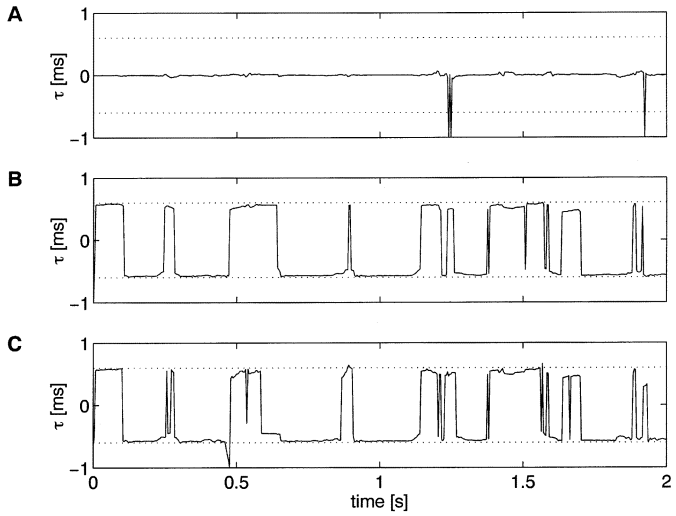

Fig. 8. Estimated inter-channel time differences (ICTDs) τ for the three synthesized stereophonic signals of Table I in the band centered at 1008 Hz.

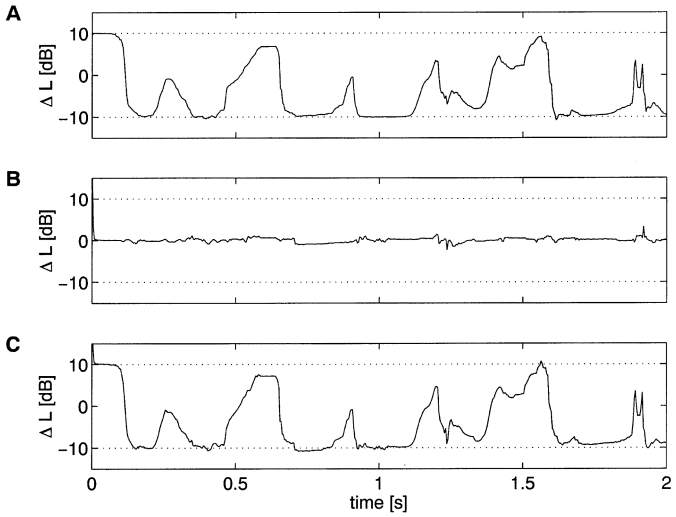

Fig. 9. Estimated inter-channel level differences (ICLDs) for the three synthesized stereophonic signals of Table I in the band centered at 1008 Hz.

The BCC-processed signals were obtained by analyzing the stereophonic input (reference) signal with the CFB-based analyzer to generate the spatial cues and by creating the mono signal. For a sampling rate of 32 kHz, ICLDs of 21 partitions were estimated and transmitted to the synthesizer every 128 samples (4 ms). The stereophonic signal was reconstructed from the mono signal by restoring the generated spatial cues with an FFT-based synthesizer. The synthesizer was operated with a size 512 FFT and an effective sine-window length of 256, zero padded to 512.

The 9 participating subjects were familiarized with the test procedure and potential artifacts using training signals.

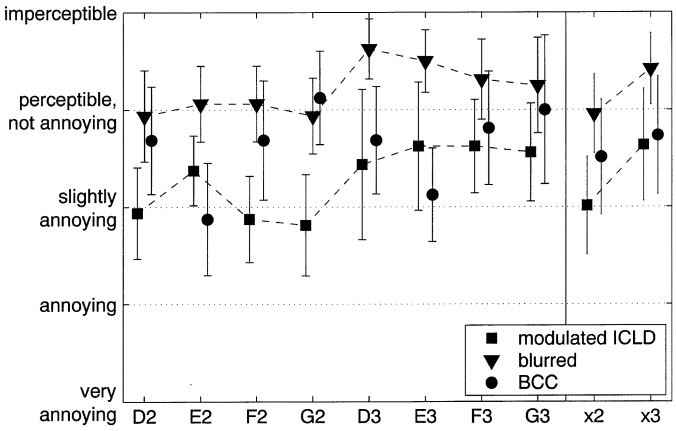

Fig. 10. Subjective grades according to the ITU-R five-grade scale and 95% confidence intervals. [Average grades of two-source mixes (x2) and three-source mixes (x3)].

After the training, each subject was asked to assess the overall quality of the anchor signals and the BCC-processed signals with respect to the reference in a listening test based on ITU-R BS.1116 [32] using the ITU-R five-grade-impairment scale. Additionally, subjects were asked to specify the kind of degradation they perceive. They were given a choice of specifying "reduced image width," "reduced image stability," "increased image blur," or "other degradations." It was permitted to choose more than one kind of degradation for each test item. The test was performed by each subject sitting at the standard listening position ("sweetspot") for conventional stereophonic loudspeaker playback. The playback room acoustics was similar to the specifications of ITU-R BS.1116, however, not fully compliant. The test items were played back from a computer under the subject's control and with comfortable volume.

Some trends can be observed from the results in Fig. 10. The quality of the anchor signals is almost equal among the two-phantom-source signals and among the three-phantom-source signals, the latter having a consistently higher perceived quality. However, the average quality of BCC appears to depend less on the number of phantom sources than it depends on the signal category.

A nonformal analysis of the subjective results concerning the perceived kind of degradation of BCC suggests that reduced image width is the dominant degradation for all BCC-processed items tested. A degradation caused by a stability reduction is in general less prominent and shows a larger variation over the different item categories. For instance, the detection probability of reduced stability for "vocals" is significantly higher than for "talkers." A degradation due to increased blur is perceived with lowest probability. Other kinds of degradations, such as sound coloration or other artifacts, were reported to be insignificant. These results show that BCC is able to reconstruct two or three phantom sources in a stereophonic signal from mono with a signal-dependent degradation of the spatial image.

B. CFB Versus FFT-Based Synthesis

A third experiment was designed to assess the impact of different FFT sizes used for the BCC synthesis on the audio quality with respect to the CFB-based reference synthesis. The excerpts were processed with the CFB-based analyzer to estimate the ICLDs. These ICLDs were resampled to match the time/frequency resolution of the FFT-based synthesis. The synthesizer introduces the ICLD in each band when generating the reconstructed stereophonic signal according to Fig. 6. The subband representation of the mono signal is computed in the synthesizer by applying a forward FFT of the same size as the inverse FFT. The weighting factors w_L and w_R are derived from the ICLDs.

TABLE II
FILTER BANK (FB) PARAMETERS USED FOR EXPERIMENT 3

label	*FB*	*size*
CFB	CFB	98 non-uniform bands
FT2k	FFT	2048
FT1k	FFT	1024
FT512	FFT	512
FT256	FFT	256

TABLE III
LIST OF THE AUDIO EXCERPTS USED IN EXPERIMENT 3. THE LAST TWO COLUMNS CONTAIN THE SOURCES OF EXCERPT 1, 2, AND 3, THAT ARE PLACED TO THE LEFT OR RIGHT SIDE OF THE SPATIAL IMAGE BY IMPOSING A LEVEL DIFFERENCE (AMPLITUDE PANNING)

excerpt	*category*	*left*	*right*
1	speech	male	female
2	singing	tenor	soprano
3	percussions	castanets	drums
4	applause	(stereo recording)	

Table II lists the five synthesizer configurations used in the test. The filter bank (FB) type is either CFB or FFT. Four different FFT sizes were used to evaluate the impact of different time/frequency resolutions on the audio quality. The parameter choices are motivated by experimental psychoacoustic data [25] that shows a binaural time/frequency resolution in the same range. Moreover, preliminary experiments suggested that FFT lengths below 256 should be excluded since the average overall audio quality is not improved but the side information rate is potentially increased. All forward and inverse FFTs use time-domain sine windows with 50% overlap.

Four different stereophonic audio excerpts, each with a duration of approximately 10 s sampled at 32 kHz, were used in the test. Table III summarizes their contents. The first three excerpts were identical to excerpt D2, E2, and G2 of the previous test. The fourth excerpt is a stereophonic recording of applause. It is known as a critical signal for joint-stereo coding since the spatial image is very dynamic.

The test items, including the reference excerpt with its differently processed versions, were presented over loudspeakers under the same acoustical conditions as in the previous test. The five participating subjects were asked to grade different specific distortions and the overall audio quality of the processed excerpts with respect to the known reference, i.e., the original excerpt. For this assessment, the testing scheme of the previous

TABLE IV
TASKS AND SCALES OF THE SUBJECTIVE TEST IN EXPERIMENT 3

task		*scale*
1	image width	stereo...mono
2	image stability	stable...unstable
3	audio quality ignoring spatial image distortions	ITU-R 5-grade impairment
4	overall audio quality	ITU-R 5-grade impairment

test was modified. The scheme allows for a more accurate measurement of image and other distortions on continuous scales as opposed to a "yes-or-no" decision. The image blur was not assessed in this test since it does not contribute significantly to the overall image distortions. No anchor signals were used. The four different grading tasks of this test are summarized in Table IV. Tasks 1 and 2 assess the two properties of the reproduced spatial image that are thought to determine the spatial image quality, i.e. width and stability. Task 3 evaluates distortions introduced by the stereophonic synthesis that do not result in image artifacts. For example, aliasing and blocking artifacts should be detected here. Task 4 is an important measure for global optimization of BCC.

During the test, each subject was able to randomly access each test item processed by the five different synthesizers and the reference by using the corresponding "Play" button of a graphical interface. This play function stops a possibly active audio output at any time, such that the subject can do quick initial listening through all items before proceeding with a more thorough evaluation. The grades were entered via graphical sliders that are permanently visible for all test items and can be adjusted at any time to reflect the proper grades and ranking. It is important to note that subjects were specifically asked to pay attention to the rank order of the test items. The feature of being able to play the items according to their rank order greatly facilitates this task as opposed to other testing schemes that allow to listen only once to each item in a pre-defined order. The ordering of the synthesizers was randomly chosen for each subject and each excerpt but not changed during the four different tasks performed for each excerpt. The philosophy of this test method corresponds closely to MUSHRA [33].

The experimental results are shown for the individual excerpts only. Averaging over the grades of different excerpts cannot be justified due to substantially deviating ratings. The grades of each task will be discussed in the following subsections.

1) Image Width: The grades for image width are shown in Fig. 11 for each excerpt and each synthesizer with respect to the reference. Apparently, all synthesizers reduce the image width for all test items. For excerpt 2 there is a trend toward a smaller image width with reduced FFT size. This trend is reverse for excerpt 3. This result can be explained by the more stationary character of excerpt 2 (singing) requiring higher frequency resolution in contrast to the nonstationary excerpt 3 (percussions) which requires a higher time resolution for a proper image reproduction. The overall performance of the 256-point FFT is similar to the CFB performance.

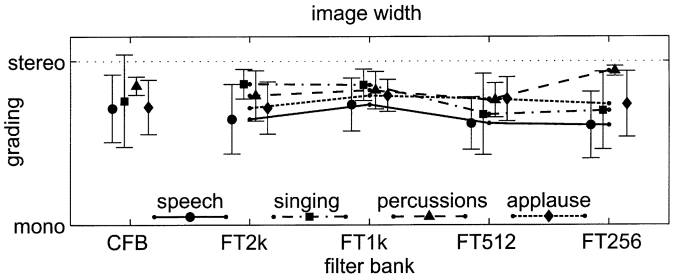

Fig. 11. Perceived image width and 95% confidence intervals (error bars are horizontally offset to increase readability).

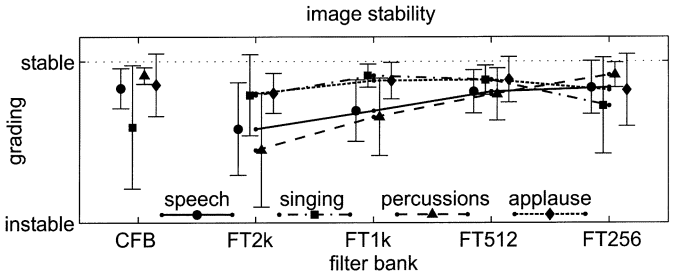

Fig. 12. Perceived image stability and 95% confidence intervals.

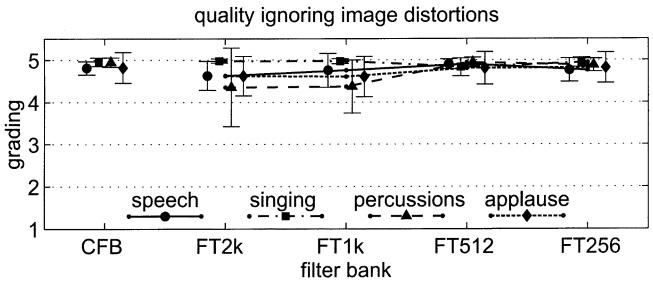

Fig. 13. Perceived audio quality ignoring spatial image distortions with 95% confidence intervals.

2) Image Stability: Grades for image stability are given in Fig. 12. The image stability is best if the virtual sound source location is stationary. Source locations are well defined for the reference excerpts 1, 2, and 3. However, for excerpt 4 (applause) each source is only active for a short time so that a moving source cannot be detected. That is why excerpt 4 appears close to "stable" for all synthesizers. From the remaining excerpts, 1 and 3 are more critical than 2. For excerpt 1 and 3, the stability increases consistently with time resolution of the FFT-based synthesizer. For excerpt 2 an FFT with medium time resolution shows best grades. The CFB-based synthesis performs equally or better than the FFT-based schemes except for excerpt 2.

3) Quality, Ignoring Image Distortions: In task 3 the audio quality is assessed without considering image degradations. The results in Fig. 13 show no significant degradations except for excerpt 3 which appears critical for the size 2048 and 1024 FFT. The time resolution is apparently insufficient for this excerpt (percussions) containing many transients.

4) Overall Quality: The overall quality grades in Fig. 14 show the integral impact of all noticeable distortions on audio

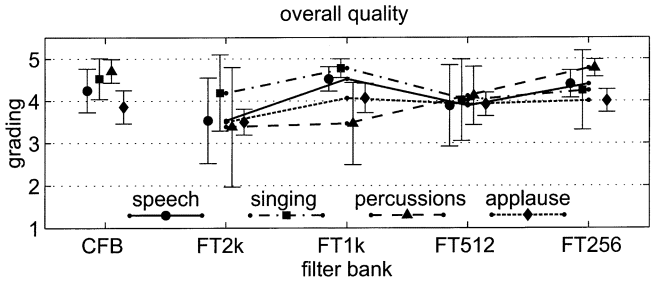

Fig. 14. Perceived audio quality and 95% confidence intervals.

quality to facilitate the selection of the synthesizer with best overall performance. Obviously, the overall quality reflects the influence of the degradations assessed in task 1, 2, and 3 and it combines these individual components into a perceptually meaningful global measure.

From visual inspection it is concluded that among the FFT-based schemes the 256-point synthesis has best performance for the test excerpts followed by the 512-point FFT. The 1024 and 2048 size FFT show significantly reduced quality for at least one excerpt. The 256-point FFT has a clear advantage over longer FFTs for excerpt 3 (percussions) which requires a high time resolution. For the more stationary excerpts 1 and 2, an FFT length between 256 and 1024 reaches about the same quality. On average the CFB-based synthesis has the same performance as the 256-size FFT. It is interesting to note, that the time resolution of the CFB and the 256-point FFT at 500 Hz is higher than the measured binaural time resolution summarized in [25]. The frequency resolution of the 256-point FFT is slightly lower than experimental data of [25].

V. Conclusions

Binaural Cue Coding (BCC) exploits auditory spatial cues for efficient data rate reduction and spatial rendering. BCC for Natural Rendering includes an analyzer for spatial cue estimation and a synthesizer for spatial cue restoration based on a down-mixed one-channel signal. The most relevant spatial cues are inter-channel level differences, time differences, and correlation. A systematic BCC design approach takes advantage of existing binaural perception models. A design example is the BCC reference scheme based on a Cochlear Filter Bank for stereophonic two-channel audio. The perceived quality of this BCC analysis/synthesis scheme using level-difference cues only is investigated for loudspeaker playback. The results show that the perceived degradation is mainly caused by a reduced auditory image width and stability. Other distortions are negligible.

A low-complexity FFT-based BCC synthesizer implementation is presented and evaluated. The best performing FFT-based BCC synthesizer has an FFT size of 256 at 32 kHz sampling rate and shows equivalent performance to the reference BCC synthesizer. This implementation is suitable for low-cost real-time systems.

Several enhancements of the FFT-based scheme are presented in Part II including multichannel audio and a BCC scheme for flexible rendering. The enhanced schemes can employ inter-channel time-difference and correlation cues in addition to level-difference cues.

Future work includes the evaluation of BCC applied to audio with more than two channels. More research is also necessary to fully understand the perceptual aspects of rendering spatial images with a BCC synthesizer.

Acknowledgment

The authors thank P. Kroon and T. Gänsler for helpful suggestions on the draft manuscript. They thank the anonymous reviewers for valuable comments.

References

[1] *Generic Coding of Moving Pictures and Associated Audio Information—Part 7: Advanced Audio Coding*, ISO/IEC Std. 13 818-7, 1997.
[2] D. Sinha, J. D. Johnston, S. Dorward, and S. R. Quackenbush, *The Digital Signal Processing Handbook*. New York: IEEE Press, 1998, ch. 42, pp. 42-1–42-18.
[3] J. D. Johnston and A. J. Ferreira, "Sum-difference stereo transform coding," in *Proc. IEEE ICASSP*, 1992, pp. 569–572.
[4] J. Herre, K. Brandenburg, and D. Lederer, "Intensity stereo coding," in *Proc. 96th AES Conv., Feb. 1994*, Amsterdam, 1994.
[5] H. Fuchs, "Improving joint stereo audio coding by adaptive inter-channel prediction," in *Proc. IEEE WASPAA*, Mohonk, NY, Oct. 1993.
[6] C. Faller and F. Baumgarte, "Efficient representation of spatial audio using perceptual parametrization," in *Proc. IEEE WASPAA*, Mohonk, NY, Oct. 2001.
[7] F. Baumgarte and C. Faller, "Estimation of auditory spatial cues for binaural cue coding," in *Proc. ICASSP 2002*, Orlando, FL, May 2002.
[8] C. Faller and F. Baumgarte, "Binaural cue coding: a novel and efficient representation of spatial audio," in *Proc. ICASSP 2002*, Orlando, FL, May 2002.
[9] F. Baumgarte and C. Faller, "Why binaural cue coding is better than intensity stereo coding," in *Proc. AES 112th Conv.*, Munich, Germany, May 2002.
[10] C. Faller and F. Baumgarte, "Binaural cue coding applied to stereo and multi-channel audio compression," in *Proc. AES 112th Conv.*, Munich, Germany, May 2002.
[11] F. Baumgarte and C. Faller, "Design and evaluation of binaural cue coding," in *AES 113th Conv.*, Los Angeles, CA, Oct. 2002.
[12] C. Faller and F. Baumgarte, "Binaural cue coding applied to audio compression with flexible rendering," in *Proc. AES 113th Conv.*, Los Angeles, CA, Oct. 2002.
[13] J. L. Hall, "Auditory psychophysics for coding applications," in *The Digital Signal Processing Handbook*, V. Madisetti and D. B. Williams, Eds. Boca Raton, FL: CRC, 1998, pp. 39-1–39-22.
[14] J. Blauert, *Spatial Hearing. The Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press, 1983.
[15] C. Faller and F. Baumgarte, "Binaural cue coding—Part II: Schemes and applications," *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 520–531, Nov. 2003.
[16] F. Rumsey, *Spatial Audio*. Oxford, U.K.: Focal, 2001.
[17] V. Pulkki and M. Karjalainen, "Localization of amplitude-panned virtual sources I: stereophonic panning," *J. Audio Eng. Soc.*, vol. 49, no. 9, pp. 739–752, Sept. 2001.
[18] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1994.
[19] F. L. Wightman and D. J. Kistler, *Binaural and Spatial Hearing in Real and Virtual Environments*. Princeton, NJ: Lawrence Erlbaum, 1997, ch. 1, pp. 1–23.
[20] E. A. Macpherson and J. C. Middlebrooks, "Listener weighting of cues for lateral angle: the duplex theory of sound localization revisited," *J. Acoust. Soc. Amer.*, vol. 111, no. 5, pp. 2219–2236, May 2002.
[21] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1648–1661, Mar. 1992.
[22] R. M. Stern and C. Trahiotis, *Binaural and Spatial Hearing in Real and Virtual Environments*. Princeton, NJ: Lawrence Erlbaum, 1997, ch. 24.
[23] J. Breebart, S. v. d. Par, and A. Kohlrausch, "Binaural processing model based on contralateral inhibition. I. Model structure," *J. Acoust. Soc. Amer.*, vol. 110, no. 2, pp. 1074–1088, Aug. 2001.
[24] M. v. d. Heijden and C. Trahiotis, "Binaural detection as a function of interaural correlation and bandwidth of masking noise: Implications for estimates of spectral resolution," *J. Acoust. Soc. Amer.*, vol. 103, no. 3, pp. 1609–1614, Mar. 1998.

[25] I. Holube, M. Kinkel, and B. Kollmeier, "Binaural and monaural auditory filter bandwidths and time constants in probe tone detection experiments," *J. Acoust. Soc. Amer.*, vol. 104, no. 4, pp. 2412–2425, Oct. 1998.
[26] A. Kohlrausch, "Auditory filter shape derived from binaural masking experiments," *J. Acoust. Soc. Amer.*, vol. 84, no. 2, pp. 573–583, Aug. 1988.
[27] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, pp. 103–138, 1990.
[28] F. Baumgarte, "Improved audio coding using a psychoacoustic model based on a Cochlear Filter Bank," *IEEE Trans. Speech Audio Processing*, vol. 10, pp. 495–503, Oct. 2002.
[29] L. Lin, W. H. Holmes, and E. Ambikairajah, "Auditory filter bank inversion," in *Proc. IEEE ISCAS 2001*, May 2001, pp. II-537–II-540.
[30] J. Princen, "The design of nonuniform modulated filterbanks," *IEEE Trans. Signal Processing*, vol. 43, pp. 2550–2560, Nov. 1995.
[31] H. S. Malvar, *Signal Processing With Lapped Transforms*. Norwood, MA: Artech House, 1992.
[32] ITU-R, Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems, 1997.
[33] ITU-R, Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems, 2001.

Frank Baumgarte received the M.S. and Ph.D. (Dr.-Ing.) degrees in electrical engineering from the University of Hannover, Germany, in 1989 and 2000, respectively. During the studies and as independent consultant he implemented real-time signal processing algorithms on a variety of DSPs including a speech coder and an MPEG-1 Layer 3 decoder. His dissertation includes a nonlinear physiological auditory model for application in audio coding.

In 1999 he joined the Acoustics and Speech Research Department, Bell Labs, Lucent Technologies, Murray Hill, NJ, where he was engaged in objective quality assessment and psychoacoustic modeling for audio coding. He became a Member of Technical Staff of the Media Signal Processing Research Department, Agere Systems, a Lucent spin-off, in 2001, focusing on advanced perceptual models for multichannel audio coding, auditory scene analysis and music synthesis. His main research interests in the area of acoustic communication include the understanding and modeling of the human auditory system physiology, psychophysics, audio and speech coding, and quality assessment.

Christof Faller received the M.S. (Ing.) degree in electrical engineering from ETH Zurich, Switzerland, in 2000. During his studies, he worked as an independent consultant for Swiss Federal Labs, applying neural networks to process parameter optimization of sputtering processes and spent one year at the Czech Technical University (CVUT), Prague.

In 2000, he became a Consultant for the Speech and Acoustics Research Department, Bell Laboratories, Lucent Technologies, Murray Hill, NJ. After one and a half year consulting, partially from Europe, he became a Member of Technical Staff, focusing on new techniques for audio coding applied to digital satellite radio broadcasting. Recently, he moved to the newly formed Media Signal Processing Research Department of Agere Systems, a Lucent spin-off. His research interests include generic signal processing, specifically audio coding, control engineering, and neural networks.

Mr. Faller won first prize in the Swiss national ABB (Asea Brown Boveri) Youth Science Contest organized in honor of the 100-year existence of ABB (formerly BBC) in 1991.